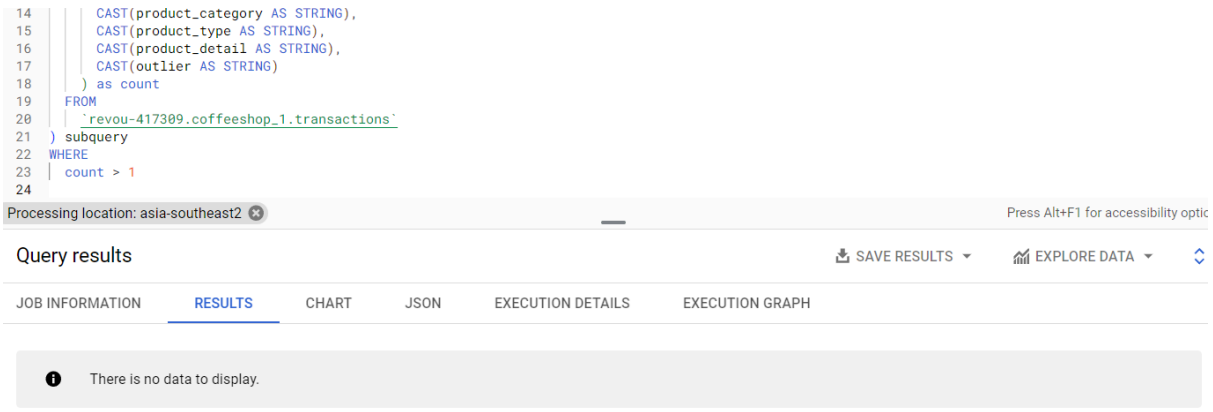
Check Data Duplicates

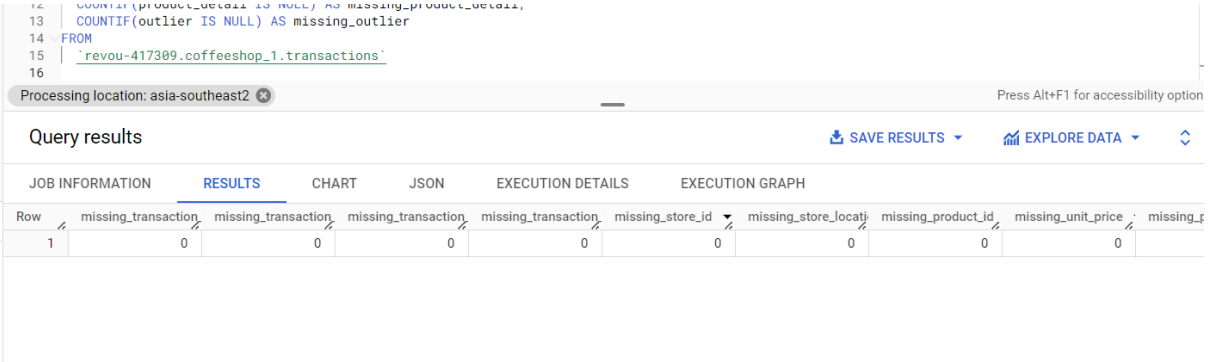
```
SELECT *
FROM (
  SELECT
    *,
    COUNT(*) OVER(PARTITION BY
      CAST(transaction_id AS STRING),
      CAST(transaction_date AS STRING),
      CAST(transaction_time AS STRING),
      CAST(transaction_qty AS STRING),
      CAST(store_id AS STRING),
      CAST(store_location AS STRING),
      CAST(product_id AS STRING),
      CAST(unit_price AS STRING),
      CAST(product_category AS STRING),
      CAST(product_type AS STRING),
      CAST(product_detail AS STRING),
      CAST(outlier AS STRING)
    ) as count
  FROM
    `revou-417309.coffeeshop_1.transactions`
) subquery
WHERE
  count > 1
```



Check For Missing Values

```
SELECT
  COUNTIF(transaction_id IS NULL) AS missing_transaction_id,
  COUNTIF(transaction_date IS NULL) AS missing_transaction_date,
  COUNTIF(transaction_time IS NULL) AS missing_transaction_time,
```

```
  COUNTIF(transaction_qty IS NULL) AS missing_transaction_qty,
  COUNTIF(store_id IS NULL) AS missing_store_id,
  COUNTIF(store_location IS NULL) AS missing_store_location,
  COUNTIF(product_id IS NULL) AS missing_product_id,
  COUNTIF(unit_price IS NULL) AS missing_unit_price,
  COUNTIF(product_category IS NULL) AS missing_product_category,
  COUNTIF(product_type IS NULL) AS missing_product_type,
  COUNTIF(product_detail IS NULL) AS missing_product_detail,
  COUNTIF(outlier IS NULL) AS missing_outlier
FROM
  `revou-417309.coffeeshop_1.transactions`
```



Check Outliers
Method 1

1. List all unique combinations of product ID and unit price

product_id	unit_price	store_id	count
1	18	8, 3, 5	209
2	18	8, 3, 5	183
3	14.75	8, 3, 5	169
4	20.45	8, 3, 5	150
5	15	8, 3, 5	148
6	21	8, 3, 5	218

…
See this.

There are different unit prices for the same product ID

2. Extract the product ID that has different unit prices, in what store and month it occurs, and how many times it occurs

product_id	unit_price	product_detail	store_id	month	count
9	23	Organic Decaf Blend	5	1, 3, 4, 6	5
9	28	Organic Decaf Blend	5	1, 2, 3, 4, 5, 6	17
9	12	Organic Decaf Blend	5	1, 4, 5, 6	7
9	22.5	Organic Decaf Blend	8, 3, 5	1, 2, 3, 4, 5, 6	177
69	3.25	Hazelnut Biscotti	8, 3, 5	1, 2, 3, 4, 5, 6	1988
69	4.06	Hazelnut Biscotti	5	1, 2, 3, 4, 5, 6	21
70	3.25	Cranberry Scone	8, 3, 5	1, 2, 3, 4, 5, 6	2053
70	4.06	Cranberry Scone	5	1, 2, 3, 4, 5, 6	24
71	3.75	Chocolate Croissant	8, 3, 5	1, 2, 3, 4, 5, 6	2819
71	4.69	Chocolate Croissant	5	1, 2, 3, 4, 5, 6	17
72	4.06	Ginger Scone	5	1, 2, 3, 5, 6	21

…
See this.

3. Analyzing the Price Fluctuations

product_id	unit_price	product_detail	store_id	month	count
9	23	Organic Decaf Blend	5	1, 3, 4, 6	5
9	28	Organic Decaf Blend	5	1, 2, 3, 4, 5, 6	17
9	12	Organic Decaf Blend	5	1, 4, 5, 6	7
9	22.5	Organic Decaf Blend	8, 3, 5	1, 2, 3, 4, 5, 6	177

Methodology

1. Data Filtering: We filtered the data to focus on specific products and prices. For example, we looked at instances where ‘product_id’ was 9 and ‘unit_price’ was 23 or 28.
2. Trend Identification: We noticed a consistent trend in the data. The price of product 9 dropped to 12 on the 7th of each month, suggesting a recurring discount or sale event. There were also instances where the price of the same product increased on specific dates, possibly due to various factors such as promotional events, dynamic pricing strategies, or other business practices.
3. Outlier Analysis: We considered the possibility that these price fluctuations could be outliers. However, given that these fluctuations were not isolated incidents and followed a consistent pattern, we concluded that they were likely part of a pricing strategy rather than errors or outliers.

Therefore, we decided to keep these instances in our analysis.

DATA ANALYSIS

1. Apa harga dapat mempengaruhi banyak sedikitnya sales product dalam product type?

Extract and group by product_type and detail

```
SELECT product_category, product_type, product_detail, unit_price, SUM(transaction_qty)
as Total_Quantity
FROM `revou-417309.coffeeshop_1.transactions_store3`
GROUP BY product_category, product_type, product_detail, unit_price
ORDER BY product_category, product_type, Total_Quantity DESC;
```

From this, we get from product_type

- Product type: Biscotti

product_type	product_detail	unit_price	Total_Quantity
Biscotti	Hazelnut Biscotti	3.25	2007
Biscotti	Chocolate Chip Biscotti	3.5	1907
Biscotti	Ginger Biscotti	3.5	1824

We see that the hazelnut Biscotty is cheaper and has more sales than the other biscotti. However, the difference is not significant.

- Product type: Pastry

Pastry	Chocolate Croissant	3.75	2839
Pastry	Croissant	3.5	1923
Pastry	Almond Croissant	3.75	1909

Here we can see that more expensive pastry has more sales. We can deduce that, the price is not relevant to the sales of pastry

- Product type: Scone

Scone	Ginger Scone	3.25	2079
Scone	Cranberry Scone	3.25	2068
Scone	Jumbo Savory Scone	3.75	2005
Scone	Scottish Cream Scone	4.5	1967
Scone	Oatmeal Scone	3	1820

- Product type: Barista Espresso

Barista Espresso	Latte	3.75	4580
Barista Espresso	Latte Rg	4.25	4488
Barista Espresso	Cappuccino	3.75	4243
Barista Espresso	Espresso shot	3	4160
Barista Espresso	Cappuccino Lg	4.25	4143
Barista Espresso	Ouro Brasileiro shot	3	2280
Barista Espresso	Ouro Brasileiro shot	2.1	959

- Product type: Drip Coffee

Drip coffee	Our Old Time Diner Blend Sm	2	4484
Drip coffee	Our Old Time Diner Blend Rg	2.5	4410
Drip coffee	Our Old Time Diner Blend Lg	3	3985

From this there is no correlation that prices affect sales for product types. Other factor could be at play
However to be sure:
Dalam analisis ini, hubungan antara harga dan penjualan produk yang berbeda di kedai kopi diperiksa. Data dikelompokkan berdasarkan product_category dan product_type, dan korelasi antara unit_price dan Total_Quantity dihitung untuk setiap kelompok.

Standar deviasi dari jumlah penjualan juga dihitung untuk menilai variabilitas dalam penjualan untuk setiap product_detail. Statistik t kemudian dihitung untuk menguji signifikansi statistik dari korelasi tersebut.

```
WITH
  product_sales AS (
  SELECT
    product_category,
    product_type,
    product_detail,
    unit_price,
    SUM(transaction_qty) as Total_Quantity
  FROM
    `revou-417309.coffeeshop_1.transactions`
```

```
  GROUP BY
    product_category,
    product_type,
    product_detail,
    unit_price
  ),
  correlation AS (
  SELECT
    product_category,
    product_type,
    corr(unit_price, Total_Quantity) as correlation_coefficient,
    COUNT(*) as n
  FROM
    product_sales
  GROUP BY
    product_category,
    product_type
  )
SELECT
  product_category,
  product_type,
  correlation_coefficient,
  (correlation_coefficient * SQRT(n - 2)) / SQRT(1 - correlation_coefficient *
correlation_coefficient) as t_statistic
FROM
  correlation;
```

Menghitung p-value

p_value = 2 * (1 - scipy.stats.t.cdf(abs(t_statistic), df=degrees_of_freedom))

Hasilnya menunjukkan bahwa tidak ada koefisien korelasi negatif yang memiliki nilai p-value kurang dari 0,1, yang menunjukkan tidak ada korelasi negatif yang signifikan antara harga dan penjualan pada tingkat signifikansi 0,1. Oleh karena itu, dapat disimpulkan bahwa harga tidak secara signifikan mempengaruhi penjualan untuk setiap jenis produk dalam setiap kategori produk, dengan asumsi asumsi analisis terpenuhi.

2. Apa ada preferensi product type setiap category product di setiap lokasi?

calculate the total sales for each product type in product category in each of the store_id (3, 5, 8)

```
SELECT
    store_id,
```

```
    product_category,
    product_type,
    SUM(transaction_qty * unit_price) as total_revenue,
    SUM(transaction_qty) as total_sales
FROM
    `revou-417309.coffeeshop_1.transactions`
WHERE
    store_id = 8
GROUP BY
    store_id, product_category, product_type
ORDER BY
    store_id,product_category, total_sales DESC
```

Hasil yang didapatkan. Dapat disimpulkan bahwa ada preferensi product type untuk setiap toko.

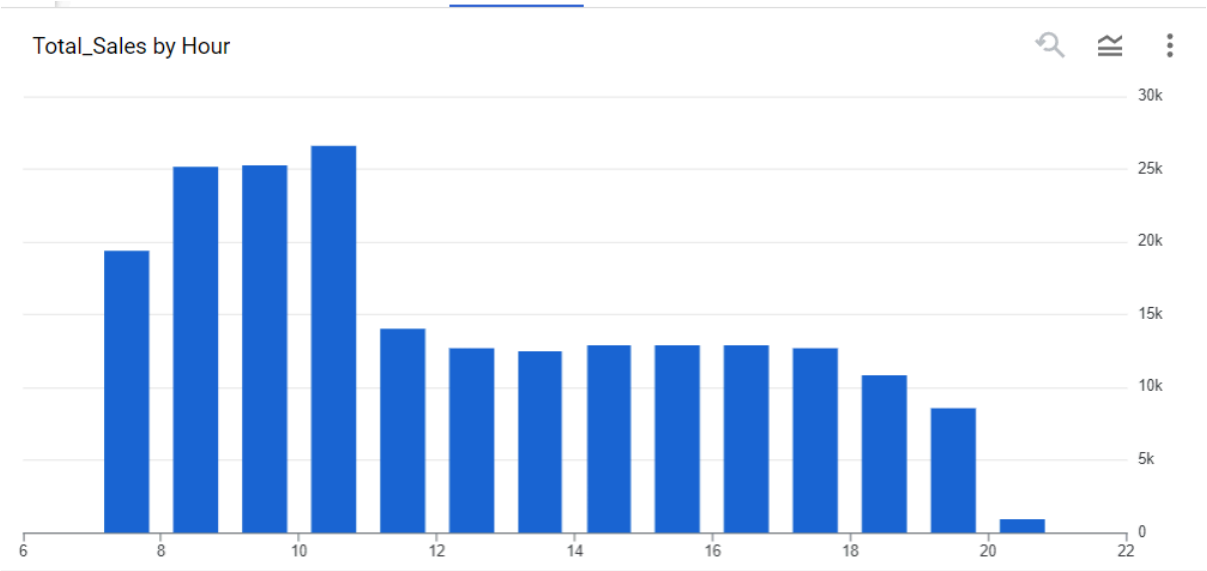
3. Apakah banyaknya variasi mempengaruhi sales product?

```
SELECT
    product_category,
    product_type,
    COUNT(DISTINCT product_detail) as num_variations,
    SUM(transaction_qty) as total_sales,
    SUM(transaction_qty * unit_price) as total_revenue
FROM
    `revou-417309.coffeeshop_1.transactions`
GROUP BY
    product_type, product_category
ORDER BY product_category, product_type
```

Dapat disimpulkan di hasilnya bahwa semakin banyak variasi, sales dan revenue semakin meningkat.

4. Apakah waktu penjualan mempengaruhi sales product cofeeshop?

```
SELECT
  EXTRACT(HOUR FROM transaction_time) as Hour,
  SUM(transaction_qty) as Total_Sales
FROM `revou-417309.coffeeshop_1.transactions`
WHERE EXTRACT(HOUR FROM transaction_time) BETWEEN 7 AND 20
GROUP BY Hour
ORDER BY Hour;
```



Dapat disimpulkan bahwa pada jam 7 - 11, adalah jam paling sibuk dan akan turun drastis pada jam 11 keatas.

Pembahasan pertanyaan Metrix :

```
WITH productcref AS (
  SELECT product_detail, SUM(transaction_qty * unit_price) AS revenue
  FROM km-revou-sql-class-416100.Coffeshop.transaksi
  GROUP BY product_detail
),
productcref2 AS (
  SELECT *, SUM(revenue) OVER () AS total_revenue
  FROM productcref
),
productcref3 AS (
  SELECT *, revenue / total_revenue AS rate
  FROM productcref2
  WHERE revenue / total_revenue < 0.1
  ORDER BY rate ASC
)
SELECT *
FROM km-revou-sql-class-416100.Coffeshop.transaksi
WHERE product_detail IN (SELECT product_detail FROM productcref3)
```